BULgarian Resource Grammar – Efficient and Robust

Petya Osenova
BulTreebank Group
Institute for Parallel Processing
Bulgarian Academy of Sciences

DELPH-IN Summit, Paris, 2 July 2010
Overview

- Matrix-based (via customization page as of January 2010)
- Full coverage of MRS test suite
- Coverage of additional language-specific phenomena
- Incorporation of full morphology using LKB types/rules
- Documentation of theory and implementation
- Preparing to link to large Bulgarian Treebank
- Ready for inclusion as DELPH-IN resource
Some Linguistic Phenomena

- Relatively free word order
- Most arguments optional
- Rich morphological agreement
- Rich clitic system
- Idiosyncrasies with auxiliaries, complementizers (of course)
# Coverage of Test Suite

## ‘1006’ Coverage Profile

<table>
<thead>
<tr>
<th>Length</th>
<th>total items</th>
<th>positive items</th>
<th>word string</th>
<th>lexical items</th>
<th>distinct analyses</th>
<th>total results</th>
<th>overall coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>#</td>
<td>φ</td>
<td>φ</td>
<td>φ</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>10 – 14</td>
<td>2</td>
<td>2</td>
<td>10.00</td>
<td>169.50</td>
<td>24.00</td>
<td>2</td>
<td>100.0</td>
</tr>
<tr>
<td>5 – 9</td>
<td>77</td>
<td>73</td>
<td>5.49</td>
<td>61.44</td>
<td>5.16</td>
<td>73</td>
<td>100.0</td>
</tr>
<tr>
<td>1 – 4</td>
<td>134</td>
<td>119</td>
<td>3.15</td>
<td>23.47</td>
<td>1.92</td>
<td>119</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>213</td>
<td>194</td>
<td>4.10</td>
<td>39.26</td>
<td>3.37</td>
<td>194</td>
<td>100.0</td>
</tr>
</tbody>
</table>

(Generated by [incr tsdb()] at 1-Jul-2010 (17:55 h))
Issues

- Revisions to Matrix types
  Semantic index of (intersective) adjectives, adverbs
  HEAD-ADJ-PHRASE not QUE 0-dlist, not POSTHEAD +
  BASIC-HEAD-OPT-COMP-PHRASE also for nominal phrases

- Desire to cache lexical rule filter
  2000+ inflectional rules
  Recomputing takes 2-3 minutes
  Changes less often than lexicon

- Cyrillic font
  Trees okay, not LKB input window
  Probably solved if use LOGON configuration
Next Steps

- Slavic language family
  Tania Avgustinova has created candidate slavic.tdl
  Russian Matrix grammar now three-level implementation
  Plan to factor Bulgarian types, to use slavic.tdl

- Bulgarian Lexicon
  Plan to extend inventory of BURGER verb lexical types

- Bulgarian Treebank
  200,000 word corpus of (partly) manually constructed trees
  Web page: bultreebank.org
  Plan to define mapping from BURGER trees to BulTreeBank

- Coverage extensions